CDC Well-Being Analysis

Contents

Data wrangling

Importing

Import the data using the SAS format.

```
library(Hmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

```
temp_zip <- tempfile()
temp_xpt <- tempfile()
download.file("https://www.cdc.gov/brfss/annual_data/2016/files/LLCP2016XPT.zip", temp_zip)
unzip(zipfile=temp_zip, exdir=temp_xpt)
import <- sasxport.get(file.path(temp_xpt, "LLCP2016.xpt "))
```

```
## Processing SAS dataset LLCP2016   ..
```

```
unlink(temp_zip)
unlink(temp_xpt)
```

List of all columns names in the data.

```
names(import)
```

```
##   [1] "x.state"   "fmonth"    "idate"     "imonth"    "iday"
##   [6] "iyear"     "dispcode"  "seqno"     "x.psu"     "ctelenm1"
##  [11] "pvtresd1"  "colghous"  "stateres"  "cellfon4"  "ladult"
##  [16] "numadult"  "nummen"    "numwomen"  "ctelnum1"  "cellfon5"
##  [21] "cadult"    "pvtresd3"  "cclghous"  "cstate1"   "landline"
```

```
##  [26] "hhadult"   "genhlth"   "physhlth"  "menthlth"  "poorhlth"
##  [31] "hlthpln1"  "persdoc2"  "medcost"   "checkup1"  "exerany2"
##  [36] "sleptim1"  "cvdinfr4"  "cvdcrhd4"  "cvdstrk3"  "asthma3"
##  [41] "asthnow"   "chcscncr"  "chcocncr"  "chccopd1"  "havarth3"
##  [46] "addepev2"  "chckidny"  "diabete3"  "diabage2"  "lastden3"
##  [51] "rmvteth3"  "sex"       "marital"   "educa"     "renthom1"
##  [56] "numhhol2"  "numphon2"  "cpdemo1"   "veteran3"  "employ1"
##  [61] "children"  "income2"   "internet"  "weight2"   "height3"
##  [66] "pregnant"  "deaf"      "blind"     "decide"    "diffwalk"
##  [71] "diffdres"  "diffalon"  "smoke100"  "smokday2"  "stopsmk2"
##  [76] "lastsmk2"  "usenow3"   "ecigaret"  "ecignow"   "alcday5"
##  [81] "avedrnk2"  "drnk3ge5"  "maxdrnks"  "flushot6"  "flshtmy2"
##  [86] "pneuvac3"  "tetanus"   "fall12mn"  "fallinj2"  "seatbelt"
##  [91] "drnkdri2"  "hadmam"    "howlong"   "hadpap2"   "lastpap2"
##  [96] "hpvtest"   "hplsttst"  "hadhyst2"  "pcpsaad2"  "pcpsadi1"
## [101] "pcpsare1"  "psatest1"  "psatime"   "pcpsars1"  "bldstool"
## [106] "lstblds3"  "hadsigm3"  "hadsgco1"  "lastsig3"  "hivtst6"
## [111] "hivtstd3"  "hivrisk4"  "pdiabtst"  "prediab1"  "insulin"
## [116] "bldsugar"  "feetchk2"  "doctdiab"  "chkhemo3"  "feetchk"
## [121] "eyeexam"   "diabeye"   "diabedu"   "painact2"  "qlmentl2"
## [126] "qlstres2"  "qlhlth2"   "medicare"  "hlthcvr1"  "delaymed"
## [131] "dlyother"  "nocov121"  "lstcovrg"  "drvisits"  "medscost"
## [136] "carercvd"  "medbill1"  "medadvic"  "undrstnd"  "written"
## [141] "caregiv1"  "crgvrel1"  "crgvlng1"  "crgvhrs1"  "crgvprb2"
## [146] "crgvpers"  "crgvhous"  "crgvmst2"  "crgvexpt"  "cimemlos"
## [151] "cdhouse"   "cdassist"  "cdhelp"    "cdsocial"  "cddiscus"
## [156] "ssbsugr2"  "ssbfrut2"  "calrinfo"  "marijana"  "usemrjna"
## [161] "asthmage"  "asattack"  "aservist"  "asdrvist"  "asrchkup"
## [166] "asactlim"  "asymptom"  "asnoslep"  "asthmed3"  "asinhalr"
## [171] "imfvplac"  "hpvadvc2"  "hpvadsht"  "shingle2"  "numburn2"
## [176] "cncrdiff"  "cncrage"   "cncrtyp1"  "csrvtrt1"  "csrvdoc1"
## [181] "csrvsum"   "csrvrtrn"  "csrvinst"  "csrvinsr"  "csrvdein"
## [186] "csrvclin"  "csrvpain"  "csrvctl1"  "profexam"  "lengexam"
## [191] "pcpsade1"  "pcdmdecn"  "sxorient"  "trnsgndr"  "rcsgendr"
## [196] "rcsrltn2"  "casthdx2"  "casthno2"  "emtsuprt"  "lsatisfy"
## [201] "qlactlm2"  "useequip"  "qstver"    "qstlang"   "mscode"
## [206] "x.ststr"   "x.strwt"   "x.rawrake" "x.wt2rake" "x.chispnc"
## [211] "x.crace1"  "x.cprace"  "x.cllcpwt" "x.dualuse" "x.dualcor"
## [216] "x.llcpwt2" "x.llcpwt"  "x.rfhlth"  "x.phys14d" "x.ment14d"
## [221] "x.hcvu651" "x.totinda" "x.michd"   "x.ltasth1" "x.casthm1"
## [226] "x.asthms1" "x.drdxar1" "x.exteth2" "x.alteth2" "x.denvst2"
## [231] "x.prace1"  "x.mrace1"  "x.hispanc" "x.race"    "x.raceg21"
## [236] "x.racegr3" "x.race.g1" "x.ageg5yr" "x.age65yr" "x.age80"
## [241] "x.age.g"   "htin4"     "htm4"      "wtkg3"     "x.bmi5"
## [246] "x.bmi5cat" "x.rfbmi5"  "x.chldcnt" "x.educag"  "x.incomg"
## [251] "x.smoker3" "x.rfsmok3" "x.ecigsts" "x.curecig" "drnkany5"
## [256] "drocdy3."  "x.rfbing5" "x.drnkwek" "x.rfdrhv5" "x.flshot6"
## [261] "x.pneumo2" "x.rfseat2" "x.rfseat3" "x.drnkdrv" "x.rfmam2y"
## [266] "x.mam5021" "x.rfpap33" "x.rfpsa21" "x.rfblds3" "x.col10yr"
## [271] "x.hfob3yr" "x.fs5yr"   "x.fobtfs"  "x.crcrec"  "x.aidtst3"
```

Select the useful columns from the data and remove any records with NA's.

```
# These are the names of the fields we are taking from the CDC data.
data = import[c('avedrnk2', 'alcday5', 'smokday2', 'x.ageg5yr', 'physhlth')]
```

```r
# Rename the columns to more useful and readable names (same order as above).
colnames(data) <- c('average_drinks_per_day_drank',
                    'days_in_last_30_had_a_drink',
                    'smoking_frequency',
                    'demographic_generation',
                    'physically_healthy_days_in_last_30')
```

Let's take a look at some simple statistics for each field.

NOTE: No preprocessing or cleaning has been performed on the data yet.

```r
summary(data)
```

```
##  average_drinks_per_day_drank days_in_last_30_had_a_drink
##  Min.   : 1.00                Min.   :101.0
##  1st Qu.: 1.00                1st Qu.:202.0
##  Median : 2.00                Median :230.0
##  Mean   : 3.56                Mean   :527.3
##  3rd Qu.: 3.00                3rd Qu.:888.0
##  Max.   :99.00                Max.   :999.0
##  NA's   :249223               NA's   :18348
##  smoking_frequency demographic_generation
##  Min.   :1.00      Min.   : 1.00
##  1st Qu.:2.00      1st Qu.: 5.00
##  Median :3.00      Median : 8.00
##  Mean   :2.43      Mean   : 7.82
##  3rd Qu.:3.00      3rd Qu.:11.00
##  Max.   :9.00      Max.   :14.00
##  NA's   :282147
##  physically_healthy_days_in_last_30
##  Min.   : 1.00
##  1st Qu.:15.00
##  Median :88.00
##  Mean   :60.72
##  3rd Qu.:88.00
##  Max.   :99.00
##  NA's   :3
```

```r
nrow(data)
```

```
## [1] 486303
```

Cleaning and preprocessing

Removing useless and empty answers

Ignore NAs in 'average_drinks_per_day_drank' as that variable encodes both missing and zero values as NA.

```r
data = data[!with(data, is.na(days_in_last_30_had_a_drink) |
                        is.na(smoking_frequency) |
                        is.na(demographic_generation) |
                        is.na(physically_healthy_days_in_last_30)), ]
```

Remove people who reported any question with "Don't know" or "Refuse to answer."

Each question encodes such answers differently:

- Typically, one of 7, 77, or 777 are used to indicate a "don't know" answer.
- 9, 99, and 9999 typically indicate "refuse" answers.

```
data = subset(data, !(average_drinks_per_day_drank %in% c(77, 99)))
data = subset(data, !(days_in_last_30_had_a_drink %in% c(777, 999)))
data = subset(data, !(smoking_frequency %in% c(7, 9)))
data = subset(data, demographic_generation != 14)
data = subset(data, !(physically_healthy_days_in_last_30 %in% c(77, 99)))
```

Physical health

An answer encoded as 88 indicates that the person has no unhealthy days. Recode such answers to 0.

The answers to this question are encoded as number of unhealthy days. To transform the answers to be number of healthy days, subtract the answers from 30.

```
data$physically_healthy_days_in_last_30[data$physically_healthy_days_in_last_30 == 88] <- 0
data$physically_healthy_days_in_last_30 = 30 - data$physically_healthy_days_in_last_30
```

Demographic generation

Remove people who are not in any of the targeted age ranges. Groups 1-10 cover ages 18-69. We're not analyzing the health of anyone above 69 years old.

```
data = subset(data, demographic_generation %in% c(1:10))
```

Recoding variables into groups

Demographic generation

Recode age groups into demographics by joining smaller ones into larger ones:

- **Groups 1-3** (ages 18-24, 25-29, 30-34): *Millenials*.
- **Groups 4-6** (ages 35-39, 40-44, 45-49): *Generation X*.
- **Groups 7-10** (ages 50-54, 54-59, 60-64, 65-69): *Baby Boomers*.

```
data$demographic_generation = cut(data$demographic_generation,
                                  breaks=c(-Inf, 3, 6, Inf),
                                  labels=c('Millenials', 'Generation X', 'Baby Boomers'))
```

Drinking frequency

Let the objective of this section be to compute a respondent's average number of drinks consumed per week.

This section is quite complicated. The data we're given does not provide a straightforward to compute the average number of drinks someone consumes per week. Fortunately, this measurement can be computed using several other variables with which we are provided:

- **Average number of drinks consumed per day that a person drinks**. This variable is encoded as blank/NA for people who have not drank at all in the last month.
- **Number of days in the last 30 that a person drank at all**. This variable is encoded strangely; respondents could either answer how many days in the last week that they drank, or the amount of days in the last month that they drank.

The intuition for what we're going to do is format each of these variables such that we can apply the equation `drinking_frequency = average_drinks_per_day_drank * days_in_last_30_had_a_drink`.

"Number of days in the last 30 that a person drank at all" is encoded somewhat strangely as follows:

1. **If the respondent answered with the number of days in the last week that they drank**, then the field takes on a value from 101-107, where subtracting 100 gives the true number of days in which the person drank.
2. **If the respondent answered with the number of days in the last month that they drank**, then the field takes on a value from 201-230, where subtracting 200 gives the true number of days in which the person drank.
3. **If the respondent did not drink at all in the last month**, then the field takes on a value of 888.

To make it easier to understand each step, let's transform this variable into the number of times in the last month that the person drank. We can do this by:

- Subtracting 100 from case 1 and then multiplying by 4 weeks in a month to extrapolate how much the respondent drank in the last month.
- Subtracting 200 from case 2.
- Recoding case 3 as 0.

Later on, you'll see that we care about the number of times in the last week that the person drank; however, we recoded the variable in the last step to days in the last 30 that the person drank because the name of the variable is "days_in_last_30_had_a_drink", so it's easier to follow the logic when naming conventions are always consistent.

```
days = data$days_in_last_30_had_a_drink

data$days_in_last_30_had_a_drink[days == 888] <- 0
data$days_in_last_30_had_a_drink[days >= 100 & days < 200] <- (data$days_in_last_30_had_a_drink[days >=
data$days_in_last_30_had_a_drink[days >= 200 & days < 300] <- data$days_in_last_30_had_a_drink[days >= 1
```

Now that we have the number of days in the last 30 in which the respondent drank at all, we can apply the formula above. Note that our ultimate goal is to get the number of days in the last week in which the person drank, so we must divide the frequency by 4.

```
data$average_drinks_per_day_drank[is.na(data$average_drinks_per_day_drank)] <- 0
data$drinking_frequency = (data$days_in_last_30_had_a_drink * data$average_drinks_per_day_drank) %/% 4
```

Now, let's partition the respondents into three groups:

1. People who don't drink at all.
2. People who drink, but would not be classified as alcoholics (1-8 drinks/week).
3. People who would be classified as alcoholics (9+ drinks/week).

NOTE: The official diagnoses are different for men and women, so we use the mean of the two terms to classify both genders. In the future, we may want to separate the classifications out for each gender.

```
data$drinking_frequency = cut(data$drinking_frequency, breaks=c(-Inf, 0, 8, Inf), labels=c('None', 'Some
```

For the final step, let's remove the intermediate columns from the original dataset; they only add noise at this point.

```
data = within(data, rm(days_in_last_30_had_a_drink, average_drinks_per_day_drank))
```

Smoking frequency

Smoking frequency is encoded as follows:

1. Every day.
2. Some days.
3. Not at all.

This ordering is not very useful for analysis, so let's re-order it as follows:

0. Not at all.
1. Some days.
2. Every day.

```
data$smoking_frequency[data$smoking_frequency == 3] <- 0
# Swap 1 and 2 because they're in a confusing order.
data$smoking_frequency[data$smoking_frequency == 1] <- 9999
data$smoking_frequency[data$smoking_frequency == 2] <- 1
data$smoking_frequency[data$smoking_frequency == 9999] <- 2
```

Next, let's rename the partitions from numbers, which are difficult to read, into the factors "None", "Some Days", and "Every Day".

```
data$smoking_frequency = cut(data$smoking_frequency, breaks=c(-Inf, 0, 1, Inf), labels=c('None', 'Some
```

Write the data to a file

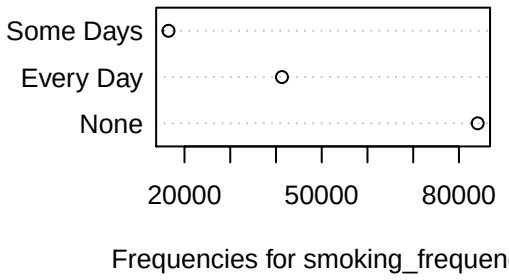
```
write_data = data[sample(nrow(data), 1000), ]
write_data$demographic_generation <- as.numeric(write_data$demographic_generation)
write_data$smoking_frequency <- as.numeric(write_data$smoking_frequency)
write_data$drinking_frequency <- as.numeric(write_data$drinking_frequency)
write.csv(write_data, "data.csv")
rm(write_data)
```

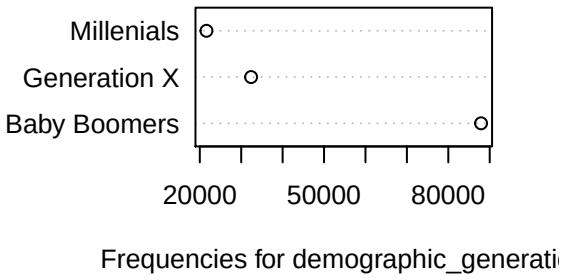
Data exploration

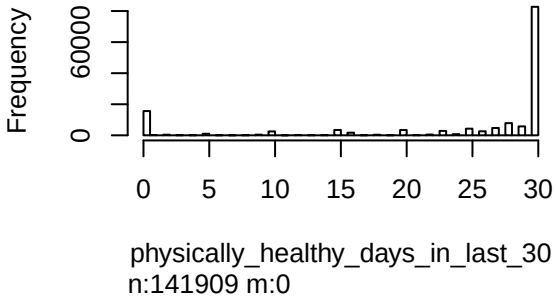
High-level overview

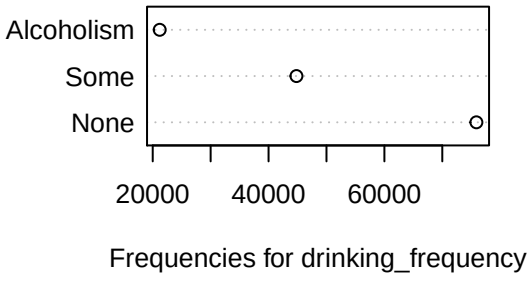
Now that we have completed all the data wrangling and preprocessing, let's explore the data for trends which we can spot visually.

```
hist(data)
```









Group means

A quick check of the healthy days group means for the categories within each predictor will tell us if there's any merit to further analysis at all.

```
aggregate(data$physically_healthy_days_in_last_30, list(data$drinking_frequency), mean)
```

```
##      Group.1        x
## 1       None 22.73613
## 2       Some 26.37747
## 3 Alcoholism 26.18156
```

```
aggregate(data$physically_healthy_days_in_last_30, list(data$smoking_frequency), mean)
```

```
##     Group.1        x
## 1      None 25.16438
## 2 Some Days 23.65168
## 3 Every Day 23.14601
```

```
aggregate(data$physically_healthy_days_in_last_30, list(data$demographic_generation), mean)
```

```
##        Group.1        x
## 1   Millenials 26.70193
## 2 Generation X 25.06583
## 3 Baby Boomers 23.59046
```

```
aggregate(data$physically_healthy_days_in_last_30, list(data$drinking_frequency, data$smoking_frequency
```

```
##       Group.1   Group.2      Group.3        x
## 1        None      None   Millenials 26.38840
## 2        Some      None   Millenials 27.88177
## 3  Alcoholism      None   Millenials 27.94716
```

```
## 4        None Some Days   Millenials 25.85934
## 5        Some Some Days   Millenials 27.07825
## 6  Alcoholism Some Days   Millenials 27.63693
## 7        None Every Day   Millenials 25.33728
## 8        Some Every Day   Millenials 26.78181
## 9  Alcoholism Every Day   Millenials 26.32345
## 10       None      None Generation X 24.61107
## 11       Some      None Generation X 27.49449
## 12 Alcoholism      None Generation X 27.50566
## 13       None Some Days Generation X 21.86509
## 14       Some Some Days Generation X 26.16578
## 15 Alcoholism Some Days Generation X 26.90733
## 16       None Every Day Generation X 22.27802
## 17       Some Every Day Generation X 25.35142
## 18 Alcoholism Every Day Generation X 25.11901
## 19       None      None Baby Boomers 22.87795
## 20       Some      None Baby Boomers 26.59996
## 21 Alcoholism      None Baby Boomers 26.74269
## 22       None Some Days Baby Boomers 19.58519
## 23       Some Some Days Baby Boomers 24.56920
## 24 Alcoholism Some Days Baby Boomers 24.03523
## 25       None Every Day Baby Boomers 20.62819
## 26       Some Every Day Baby Boomers 23.94534
## 27 Alcoholism Every Day Baby Boomers 23.82686
```

From the results above, it's impossible to say whether there is an effect in each group or not; however, with our large sample size, and the substantial differences between group means, we can speculate that further analysis is promising.

```
fit <- aov(physically_healthy_days_in_last_30 ~ drinking_frequency * smoking_frequency * demographic_ge
summary(fit)
```

```
##                                                                  Df
## drinking_frequency                                                2
## smoking_frequency                                                 2
## demographic_generation                                            2
## drinking_frequency:smoking_frequency                              4
## drinking_frequency:demographic_generation                         4
## smoking_frequency:demographic_generation                          4
## drinking_frequency:smoking_frequency:demographic_generation       8
## Residuals                                                    141882
##                                                               Sum Sq
## drinking_frequency                                            452645
## smoking_frequency                                             110206
## demographic_generation                                        179195
## drinking_frequency:smoking_frequency                            5040
## drinking_frequency:demographic_generation                      24231
## smoking_frequency:demographic_generation                       10979
## drinking_frequency:smoking_frequency:demographic_generation     2325
## Residuals                                                   13353044
##                                                              Mean Sq
## drinking_frequency                                           226322
## smoking_frequency                                             55103
## demographic_generation                                        89597
## drinking_frequency:smoking_frequency                           1260
```

```
## drinking_frequency:demographic_generation                     6058
## smoking_frequency:demographic_generation                      2745
## drinking_frequency:smoking_frequency:demographic_generation    291
## Residuals                                                       94
##                                                             F value
## drinking_frequency                                        2404.775
## smoking_frequency                                          585.495
## demographic_generation                                     952.012
## drinking_frequency:smoking_frequency                        13.389
## drinking_frequency:demographic_generation                   64.366
## smoking_frequency:demographic_generation                    29.164
## drinking_frequency:smoking_frequency:demographic_generation  3.089
## Residuals
##                                                              Pr(>F)
## drinking_frequency                                         < 2e-16 ***
## smoking_frequency                                          < 2e-16 ***
## demographic_generation                                     < 2e-16 ***
## drinking_frequency:smoking_frequency                      6.55e-11 ***
## drinking_frequency:demographic_generation                  < 2e-16 ***
## smoking_frequency:demographic_generation                   < 2e-16 ***
## drinking_frequency:smoking_frequency:demographic_generation  0.00174 **
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
drop1(fit, ~., test="F")
```

```
## Single term deletions
##
## Model:
## physically_healthy_days_in_last_30 ~ drinking_frequency * smoking_frequency *
##     demographic_generation
##                                                             Df Sum of Sq
## <none>
## drinking_frequency                                           2      5100
## smoking_frequency                                            2      2101
## demographic_generation                                       2     54924
## drinking_frequency:smoking_frequency                         4       489
## drinking_frequency:demographic_generation                    4     10890
## smoking_frequency:demographic_generation                     4      7333
## drinking_frequency:smoking_frequency:demographic_generation  8      2325
##                                                                    RSS
## <none>                                                      13353044
## drinking_frequency                                          13358144
## smoking_frequency                                           13355145
## demographic_generation                                      13407968
## drinking_frequency:smoking_frequency                        13353533
## drinking_frequency:demographic_generation                   13363934
## smoking_frequency:demographic_generation                    13360377
## drinking_frequency:smoking_frequency:demographic_generation 13355370
##                                                                AIC
## <none>                                                      644933
## drinking_frequency                                          644983
## smoking_frequency                                           644951
## demographic_generation                                      645512
```

```
## drinking_frequency:smoking_frequency                          644930
## drinking_frequency:demographic_generation                     645041
## smoking_frequency:demographic_generation                      645003
## drinking_frequency:smoking_frequency:demographic_generation 644942
##                                                                F value
## <none>
## drinking_frequency                                              27.0928
## smoking_frequency                                               11.1624
## demographic_generation                                         291.7945
## drinking_frequency:smoking_frequency                             1.2986
## drinking_frequency:demographic_generation                       28.9280
## smoking_frequency:demographic_generation                        19.4784
## drinking_frequency:smoking_frequency:demographic_generation      3.0887
##                                                                  Pr(>F)
## <none>
## drinking_frequency                                           1.722e-12 ***
## smoking_frequency                                            1.421e-05 ***
## demographic_generation                                       < 2.2e-16 ***
## drinking_frequency:smoking_frequency                          0.267932
## drinking_frequency:demographic_generation                    < 2.2e-16 ***
## smoking_frequency:demographic_generation                     4.867e-16 ***
## drinking_frequency:smoking_frequency:demographic_generation  0.001742 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```